

DOCUMENT RESUME

ED 220 514

TM 820 531

AUTHOR Yen, Wendy M.
TITLE Use of Three-Parameter Item Response Theory in the Development of CTBS, Form U, and TCS.
PUB DATE Mar 82
NOTE 62p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, March 20-22, 1982).
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Achievement Tests; Bayesian Statistics; Cognitive Processes; *Data Collection; *Item Analysis; *Latent Trait Theory; Maximum Likelihood Statistics; Scoring; *Standardized Tests; *Test Construction; Test Items
IDENTIFIERS *Comprehensive Tests of Basic Skills; *Test of Cognitive Skills; Three Parameter Model

ABSTRACT

The three-parameter logistic model discussed was used by CTB/McGraw-Hill in the development of the Comprehensive Tests of Basic Skills, Form U (CTBS/U) and the Test of Cognitive Skills (TCS), published in the fall of 1981. The development, standardization, and scoring of the tests are described, particularly as these procedures were influenced by the use of item response theory (IRT). Aspects of the development of the tests described include: (1) item tryout data collection design; (2) item statistics; (3) item selection criteria and procedures; (4) data collection and analysis design for equating levels of final tests; (5) maximum likelihood scoring procedures for obtaining trait estimates from number-right scores, as well as from item-response vectors; and (6) Bayesian/IRT procedure for obtaining objective mastery scores. The usefulness of the IRT procedures in clarifying test development and standardization is discussed.
(Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED220514

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

Use of Three-Parameter Item
Response Theory
in the Development of
CTBS, Form U, and TCS

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

W. M. Yen

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Wendy M. Yen

CTB/McGraw-Hill

Presented at the National Council on Measurement
in Education meetings, New York, March 1982

TM 820 531

Abstract

The three-parameter logistic model was used by CTB/McGraw-Hill in the development of the Comprehensive Tests of Basic Skills, Form U (CTBS/U) and the Test of Cognitive Skills (TCS), published fall 1981. This paper describes the development, standardization, and scoring of the tests, particularly as these procedures were influenced by the use of item response theory (IRT). The following aspects of the development of these tests are described:

1. item tryout data collection design;
2. item statistics;
3. item selection criteria and procedures;
4. data collection and analysis design for equating levels of final tests;
5. maximum likelihood scoring procedures for obtaining trait estimates from number-right scores, as well as from item-response vectors; and
6. Bayesian/IRT procedure for obtaining objective mastery scores.

The usefulness of the IRT procedures in the test development and standardization is evaluated.

Use of Three-Parameter Item Response Theory in the
Development of CTBS, Form U, and TCS

Overview

The Comprehensive Tests of Basic Skills, Form U (CTB/McGraw-Hill, 1981) is an achievement battery containing tests in reading, language, mathematics, reference skills, science, and social studies. Each test is designed to measure two or more category objectives. The nine levels of the battery span grades K-12. Table 1 shows the number of items by test and level for the final published tests. The Test of Cognitive Skills (CTB/

Insert Table 1 about here

McGraw-Hill, 1981) has four tests - Sequences, Analogies, Memory, and Verbal Reasoning - at five levels, spanning grades 2-12. There are 20 items for each test and level except Verbal Reasoning Level 1 which has 19 items. Tryout data were collected for both batteries in the fall of 1979. Statistical information about the items was examined by editorial staff who selected the final tests. TCS was standardized in the fall of 1980 and CTBS in the fall of 1980 and the spring of 1981. The tests were published and scoring reports made available in the fall of 1981. This paper focuses on the IRT aspects of the test development and standardization. Information about other important aspects, such as the procedures followed to ensure appropriate content coverage of the tests, may be

obtained from the TCS Test Coordinator's Handbook and Guide to Interpretation, Preliminary Edition (1981), the TCS technical bulletins, and CTBS/U Test Coordinator's Handbook and technical bulletins.

Item Tryouts

Sample

The Standard item tryout sample, composed of students in public and private schools across the country, was structured to be heterogeneous. For example, items thought to be appropriate for grade 5 were tried out at grades 4, 5, and 6. For CTBS/U there was a Standard sample and a special Black sample, obtained from schools with predominantly Black enrollments. For TCS there was a Standard sample, a Black sample, and an additional Hispanic sample. Enough students were tested to obtain a target sample of at least 400 Standard, 200 Black and (for TCS) 200 Hispanic students with usable answer documents for each tryout book. The test levels for the Black and Hispanic samples were shifted up one grade relative to the Standard sample; for example, items tried out at grades 4, 5 and 6 for the Standard sample were tried out at grades 5, 6 and 7 for the Black and Hispanic samples. This shift in the samples was made so that meaningful responses to the items could be obtained from the minority groups, which historically have obtained lower scores than the majority group.

Comparisons of item performance across ethnic groups were to be made. For these comparisons the "Black" group was taken to be students in the Black, Hispanic, or Standard samples who identified themselves as Black. The "Hispanic" group (for TCS) was analogously identified. The "Other" group was taken to be students not identified as Black or Hispanic.

Testing Materials and Procedures

Two to three times as many items were tried out as would be needed for the final tests. Students took a test book containing a subset of the tryout items and an anchor test. The tryout books contained items grouped by content areas, and sometimes two or more related content areas (e.g., Reading Vocabulary and Reading Word Attack) were included in one book. The anchor tests were the Reading Vocabulary and Mathematics Computation tests from the California Achievement Tests, Form C (CAT/C, CTB/McGraw-Hill, 1977). Items tried out in the mathematics area were accompanied by the CAT/C Mathematics Computation anchor; all other tryout items were accompanied by the CAT/C Reading Vocabulary anchor. The testing was conducted by teachers.

IRT Parameter Estimates

A modified three-parameter logistic model was used. The three-parameter model (Allen & Yen, 1979; Lord, 1980) assumes that

$$P_1(\theta) = c_1 + \frac{1 - c_1}{1 + e^{-Da_1(\theta - b_1)}},$$

where $P_1(\theta)$ is the probability that an examinee with trait value θ will get item 1 correct. The item parameters are: a_1 , the item discriminating power; b_1 , the item difficulty; and c_1 , the lower asymptote or pseudo-chance-success level. D is a constant usually set at 1.7. Because of small sample sizes, the c_1 values were not estimated but were fixed at $1/A_1 - .05$, where A_1 was the number of answer choices for item 1. Item responses were scored as right, wrong, omitted, or not reached.

A modified version of LOGIST (Wood, Wingersky, & Lord, 1976) was used to obtain item parameter estimates. The modifications streamlined the program by eliminating estimation options that were not needed. LOGIST was run separately on each test in each tryout book. Examinees who did not answer at least a third of the items or who had zero or perfect scores were excluded from the LOGIST analyses. In LOGIST, omitted items are given partial credit and not reached items are treated as if they did not appear in the test for that examinee.

Bias Analyses

Analyses were performed to compare performance of the Black and Other groups on the tryout items for CTBS/U and TCS. Analogous comparisons of Hispanic and Other groups were also made for TCS. To avoid repetition, only the Black and Other comparisons are described here.

The bias analyses essentially followed the procedures proposed by Lord (1980, pp. 217-220). Comparisons across ethnic groups were performed only for the $\hat{b}_1$ values; sample sizes were considered too small to make meaningful $\hat{a}_1$ comparisons. For each test in a tryout book LOGIST was run separately on the Black and on the Other response vectors. Excluding those items with $|\hat{b}_1| > 3$ (which tend to have large standard errors), the means and standard deviations of the $\hat{b}_1$ values were set equal for the two ethnic groups, and appropriate corresponding linear transformations were made on the $\hat{a}_1$ values and $\hat{\theta}$ values in the two groups. ("Standardizing on the b_1 ," Lord, 1980, p.217). Lord's chi-square statistic (1980, p.223) was calculated comparing the $\hat{b}_1$ values for the ethnic groups. Rules were established for deciding when sample sizes were too small or when $\hat{b}_1$ values were too extreme for the chi-square test to be useful. When these conditions were identified, a No Data bias rating was given to that item. Based on the size of the chi-square value and the direction of the difference in $\hat{b}_1$ values, each tryout item was rated Good (i.e., little bias), Fair, Poor, or Very Poor with respect to bias.

For TCS an additional bias rating was obtained for each item based on a comparison of the Hispanic and the Other groups. For each TCS item a composite bias rating was produced that was in effect the worst of the Black-Other and the Hispanic-Other ratings.

Final Tryout Item Parameter Estimates

Final tryout item parameter and trait (θ) estimates were obtained using the modified LOGIST program for the combined Black and Other (CTBS/U) or the Black, Hispanic, and Other (TCS) samples. If an item had a bias rating of Poor or Very Poor for any ethnic group comparison, that ethnic group's data for that item were excluded from the analysis by means of the LOGIST "not reached" option. In such a case, the item parameters reflected the majority group performance.

If an item had $|\hat{b}_1| > 3$, the data were deemed inadequate for accurately estimating that item's parameters. In such case, the item parameters for that item were not used in subsequent analyses.

Model Fit

Because a model was being used to describe item performance, it was important to gauge the appropriateness of the model. Using the final tryout item parameter and trait estimates, a comparison of model predictions and observed item performance was obtained. Examinees were rank ordered on the basis of their $\hat{\theta}$ values, with examinees excluded if $|\hat{\theta}| > 3$. Examinees were sorted into 10 cells with 10 per cent of the sample in each cell. For each item the number of examinees in cell k who answered item i , N_{ik} , and the number of examinees in cell k who answered item i correctly, R_{ik} , was determined. The observed proportion in cell k

passing item 1 was $O_{1k} = R_{1k}/N_{1k}$. The mean trait value for the examinees in cell k who answered item 1, $\bar{\theta}_{1k}$, was determined. Considering $\bar{\theta}_{1k}$ as the independent variable and O_{1k} as the dependent variable, a fourth-degree polynomial was fitted to the observed data. Let $\hat{O}_{1k}$ be the value of the polynomial for $\bar{\theta}_{1k}$. The fit index for item 1 was

$$F_{1k} = \sum_{k=1}^{10} |\hat{O}_{1k} - E_{1k}|,$$

with

$$E_{1k} = \frac{1}{N_{1k}} \sum_{j \in \text{cell } k}^{N_{1k}} \hat{P}_1(\hat{\theta}_j),$$

and

$$\hat{P}_1(\hat{\theta}_j) = \hat{c}_1 + \frac{1 - \hat{c}_1}{1 + e^{-1.7\hat{a}_1(\hat{\theta}_j - \hat{b}_1)}}.$$

In calculating F_{1k} , $\hat{O}_{1k}$ was used in place of O_{1k} because our unpublished research had shown that with smaller sample sizes the fit statistic is more stable and more valid (i.e., more highly correlated with large sample fit statistics) when $\hat{O}_{1k}$ is used. Depending on the value of F_{1k} , an item was given a Good, Fair, or Poor fit rating.

Anchor Test

Normative data from 1977 were available for the CAT/C anchor test. Because performance on the anchor test was related to performance on the tryout items, national performance on the tryout items could be estimated. Using the relationship between CAT/C scale scores and national percentiles, target ranges for item difficulties (b_1 values) for each test being built were expressed in CAT/C scale score units. Two types of ranges were

established: nominal and proper ranges. The nominal range of item difficulties was obtained from the nominal range of the test level. The nominal range was defined to ensure an adequate degree of measurement accuracy over the span of grades for which a given test level was designed. For example, a test level may be designed for use in grades 5.0 to 5.9. The nominal range for such a test would go from the scale score at the 25th percentile of the first month of the fifth-grade school year to the scale score at the 75th percentile of the seventh month of that school year. The nominal ranges for adjacent levels overlapped. The nonoverlapping portions of the nominal ranges were defined to be the proper ranges. A larger quota of item difficulties was specified for the proper range to ensure that with functional level testing, a high level of measurement accuracy could be obtained throughout the K-12 grade range of the test battery. Figure 1 shows the nominal and proper ranges for TCS.

Insert Figure 1 about here

In a further refinement of the desired distributions of item difficulties, the nominal and proper ranges were divided into six sub-ranges, and a percentage of items desired for each subrange was specified.

Adjusted Item Difficulties and Discriminations

In order to relate the tryout item difficulties to the nominal and proper ranges, it was necessary to transform the LOGIST item difficulties to CAT/C scale score units. LOGIST produces a trait ($\hat{\theta}$) scale with a mean of 0 and standard deviation of 1 for examinees with $|\hat{\theta}| < 3$. The CAT/C scale score mean, $\bar{S}$, and standard deviation, S_S , were obtained for these examinees. The LOGIST item parameter estimates were transformed to scale score units using the following equations:

$$\hat{b}'_1 = \hat{b}_1 S_S + \bar{S}$$

$$\hat{a}'_1 = \hat{a}_1 / S_S$$

$$\hat{c}'_1 = \hat{c}_1$$

When examinees can get items correct through guessing and the three-parameter model is appropriate, an item provides most of its information at a value higher than the item difficulty (Lord, 1980, p.152), that is, at

$$\hat{b}^*_1 = \hat{b}'_1 + \frac{1}{1.7\hat{a}'_1} \ln \left[\frac{1 + \sqrt{1 + 8\hat{c}'_1}}{2} \right].$$

The values of $\hat{b}^*_1$ were compared with the subranges to determine in which subrange an item contributed the most information. Also, $\hat{b}^*_1$ was reported to editorial staff and described as the scale score at which the item measured best.

The $\hat{a}_1'$ values were small, fractional values. To increase their interpretability by non-technical staff, the following transformed item discrimination was used:

$$\begin{aligned}\hat{a}_1^* &= \hat{a}_1' \sigma(\bar{S}) \\ &= \hat{a}_1' \frac{\sigma(\bar{S})}{S_S},\end{aligned}$$

where $\sigma(\bar{S})$ was the national standard deviation in scale score units expected for a group with a mean scale score value of $\bar{S}$. $\sigma(\bar{S})$ was obtained by linear interpolation from the CAT/C norms tables. In other words, a small adjustment in the LOGIST item discriminating power was made according to whether the tryout sample was more or less heterogeneous than a national sample was expected to be. (The LOGIST discriminating power, because it is always reported relative to a trait scale with a standard deviation of 1, is influenced by group heterogeneity.) This adjustment was made for reporting the item discrimination to editors. The value of $\hat{a}_1^*$ was used in assigning an item a Good, Fair, or Poor discrimination rating.

Estimated National P-Values

Although the $\hat{b}_1'$ and $\hat{b}_1^*$ values related an item's difficulty directly to a familiar normed scale, it was also desired to have estimated national p-values (proportions passing the items) for each tryout item. To obtain

these p-values, the mean (μ_j) and standard deviation (σ_j) of the CAT/C scale scores at the $j = 1, \dots, J$ grades of interest were used. As part of the CAT/C standardization procedure, these scale scores had been normalized by grade during the Thurstonian scaling process. The estimated national p-value of tryout item i at grade j was estimated to be

$$\hat{p}_{ij} = \frac{1}{19} \sum_{k=1}^{19} \hat{p}_{ijk}(s_{jk}),$$

where

$$\hat{p}_{ijk}(s_{jk}) = \hat{c}'_i + \frac{1 - \hat{c}'_i}{1 + e^{-1.7\hat{a}'_i(s_{jk} - \hat{b}'_i)}},$$

$$s_{jk} = \mu_j + z_k \sigma_j,$$

and z_k was the k -th standard normal score corresponding to percentiles of 5, 10, ..., 95 in the normal distribution. Our unpublished research had shown that this procedure produces accurate estimates of national p-values.

Distractor Information

For each item, statistical information about the performance of all the answer choices was produced by grade. If an item's bias rating was Good or Fair for a particular ethnic group, data for that ethnic group were pooled with data for the Other group. If an item's bias rating was Poor or Very Poor, that ethnic group's data were not pooled and the answer choice statistics were calculated separately by ethnic group. Because of

the small sample sizes, when an ethnic group's statistics were calculated separately, they were not calculated by grade but for the whole ethnic group taking that item.

Three answer choice statistics were reported to the editorial staff: (1) proportion of examinees choosing each answer choice; (2) the point-biserial correlation between the answer choice and the number-right score on the rest of the test; and (3) the mean number-right score of examinees choosing that answer choice.

Overall Quality Index

To facilitate evaluation of the overall quality of each item, an overall quality index was produced as a function of the discrimination rating, fit rating, and bias rating. Table 2 shows how the individual ratings were combined into the overall quality index.

Insert Table 2 about here

Choosing Items for Tests

There were two types of tests to be chosen: a shelf edition and a linkage book. The shelf tests were the TCS and CTBS/U tests to be published.

The linkage books were to be used only in the CTBS/U standardization for equating test levels. All these tests were to be chosen by editorial staff. The editorial staff's expertise was in content, although most of them had extensive experience choosing tests using traditional statistics. The psychometric staff provided training in the meaning and use of the new IRT statistics, provided guidelines for selecting statistically ideal tests, reviewed the statistical quality of the chosen tests, and occasionally offered suggestions for alternative item selections that would improve the statistical quality of the tests.

The statistical goals of the development of the shelf tests were (1) to minimize the standard error of measurement in the nominal, and particularly in the proper, range of a test; (2) to maximize fit; (3) to minimize bias; and (4) to choose items whose answer choices performed in a reasonable manner. Minimization of the standard error curve was approximated by choosing the target percentages of items in the six sub-ranges and choosing items with maximum discrimination. The standard error of measurement was calculated as a function of scale score using the square root of the reciprocal of the test information (Lord, 1980, p.71).

The items included in the linkage books were chosen from among the final shelf CTBS/U items. For linking Levels F through J the linkage books had as many items as the shelf books; half the linkage items came from the lower of the two levels being linked and half came from the upper of the two levels being linked. For linking Levels A through F, the linkage book had at least 20 items per test (10 items from each of

the two levels being linked) and could have more than 20 items. Where possible, the proportions of items in the linkage book for two test levels reflected the proportions of items in the shelf books for the two levels being linked.

Desirable ranges for the adjusted item difficulties for the linkage items were established. These ranges fell roughly where nominal ranges for the levels being linked overlapped. The items in a linkage book had difficulties (roughly) evenly distributed across the desired range. For each level contributing to a linkage book, items were chosen to have as broad a spread of item difficulties as possible while keeping within the desirable ranges. Items with poor discrimination ratings were avoided. Items were chosen to have roughly the same proportional representation of category objectives as the shelf tests.

A great deal of statistical information had to be reviewed by the editorial staff. This information was organized on three types of computer output: Item Card printouts, Automatic Test Selection printouts, and (computer) Interactive Displays and printouts.

Item Card Printouts

An Item Card printout contained the following information for each tryout item:

1. Item identification information such as test, item sequence number, tryout book, and category objective code was provided. Each item had been tried out at a test level in which it appeared most appropriate, and this nominal test level was indicated. It was possible for an item to be used in a test level adjacent to the nominal level if the item's category objective and format were suitable for that level. The eligibility of an item to be used in an adjacent level was indicated.

2. IRT statistics were provided, including data adequacy, overall quality, bias rating(s), fit index, fit rating, discrimination ($\hat{a}_1^*$), discrimination rating, and adjusted item difficulty in scale score units ($\hat{b}_1^*$). The subrange for the nominal level in which the adjusted item difficulty fell was indicated.

3. Estimated national p-values for the fall and spring were provided for a range of grades for which the item would be appropriate.

4. Answer choice statistics were displayed by grades and, where appropriate, by ethnic group.

Automatic Test Selection Printouts

To provide a starting place for item selection, a statistically "ideal" test was selected by a computer program. The program selected items in the desired subranges, maximized the overall qualities of the selected items,

and maximized discrimination among items with the same overall quality. Editorial requirements for a minimum number of items for each category objective were also met. All of these variables were simultaneously considered using a linear programming procedure (Algorithm 449, CACM, 1971).

The Automatic Test Selection printouts contained a listing of the eligible and selected items and a statistical description of the test composed of the selected items. This statistical description focused mainly on the standard error curve of the test and predictions of national performance on the selected test (e.g., mean proportion-correct scores) by grade.

Interactive Displays and Printouts

Using the information from the Item Card and Automatic Test Selection printouts, the editorial staff refined the selected test, focusing on content considerations in particular. In refining the test, each editor employed a computerized interactive test selection program. A computer terminal displayed identification information for the items eligible for a test and indicated the items previously selected for the test. The editor replaced items as desired and received immediate feedback

on the terminal about the statistical properties of the revised test. The same test statistics were made available as were provided on the Automatic Test Selection printout. In addition the standard error curves for up to four tests could be displayed simultaneously and compared.

Item Revisions

Of course items selected for the final tests were revised on occasion because of problems identified in the tryout. However, an effort was made to minimize revisions, because tryout data on a revised item could not be relied upon to predict national performance on the item.

Standardization

Data Collection

CTBS/U was standardized or normed in the fall, 1980 and in the spring, 1981. The standardization for TCS took place in the fall, 1980. Nationally representative samples were drawn for the standardization. The obtained sample sizes are shown in Tables 3 and 4. Details of the sampling pro-

Insert Tables 3 and 4 about here

cedures can be obtained in the CTBS and TCS Test Coordinators' Handbooks; both contain short descriptions of the sampling procedures.

Estimation of IRT Parameters

More data were available from the standardization than were needed for accurate item parameter estimation. For TCS random subsamples of 1001 examinees per grade per level were drawn from the TCS standardization sample. For each level these samples were combined over grades resulting in sample sizes of 2002 to 4004 for the TCS parameter estimation. For CTBS/U random subsamples of 5000 were chosen from those taking each test level and used for parameter estimation. Our unpublished research had shown that such sample sizes were more than adequate for obtaining very stable item parameter estimates.

Item parameter estimates were to be based on LOGIST, but LOGIST is restricted to dealing with tests composed of items that all have the same number of answer choices. A second modified version of LOGIST was created that allowed simultaneous estimation of parameters for items with from two to five answer choices. The LOGIST default procedure for estimating item parameters, including the c_1 values, was used. Omits were treated as wrong answers, and not reached items were treated as "not reaches" (Wood, Wingersky, & Lord, 1976).

Interlevel Equating

Different procedures were used to perform the TCS and CTBS/U inter-level equatings.

TCS equating. For grades where TCS levels overlapped, classrooms were randomly assigned one of the two test levels to be equated. The means and standard deviations of the trait estimates of these matched samples were equated for the adjacent levels.

Maximum likelihood trait estimates were not available for examinees with zero or perfect scores. When tests of different difficulties are administered to equivalent samples, systematically different proportions of examinees are expected to have zero and perfect scores on the two tests. These examinees with zero and perfect scores must be included if the equating is not to be biased. It was felt that useful scores could be determined for these examinees based on a rational, but admittedly arbitrary, procedure. Also, while maximum likelihood trait estimates are available for examinees with extreme but not zero or perfect scores, occasionally these trait estimates are so extreme as to have little value. These extreme scores have standard errors of measurement that are very large, and differences between these extreme values have little meaning. It appeared desirable to make these scores less extreme while affecting as few scores as possible. For these reasons, floor and ceiling trait estimates were established, and the TCS equating procedure followed these steps:

1. The floor and ceiling trait values of each level were set at the ~~trait values that had standard errors of measurement ten times the minimum~~ standard error for that level. Examinees with zero or perfect scores had trait values set at the floor or ceiling. Any maximum likelihood trait estimates that were more extreme than the floor or ceiling were set equal

to the floor or ceiling.

2. The means and standard deviations of trait values for adjacent test levels for the matched samples were equated successively until all levels were placed on the same scale.

3. The average over levels of the minimum standard errors of measurement was obtained.

4. The floor and ceiling of each level were reset at the trait values that had standard errors of measurement five times the average minimum standard error found in step 3.

5. Step 2 was repeated.

CTBS/U equating. A series of linkage books containing items from each of two adjacent test levels was used in the CTBS/U interlevel equating. The numbers of items in these books are given in Table 5. To produce the CTBS/U interlevel equating, successive pairs of shelf and linkage books

Insert Table 5 about here

were analyzed together using LOGIST, pooling data for the items in common and using LOGIST's "not reached" option for the items unique to a particular

book (see Table 6). After these LOGIST runs were completed, it was necessary

Insert Table 6 about here

to link the items from the separate runs; for example, the A-B items from runs 1 and 2. This linkage was made by equating the means and standard deviations of the b_1^A values from the two runs, after eliminating any outliers found by visual inspection of item parameter plots.

To set the floors and ceilings for trait estimates, the average minimum standard error taken over the linked test levels was obtained. Also, the distribution of trait estimates for examinees who took the shelf test was obtained for each level. The floor of a test level was set at the maximum of (1) the lowest trait value with a standard error less than or equal to ten times the average minimum standard error, and (2) the lowest trait value obtained in the distribution. The ceiling or maximum score for an examinee without a perfect score was set at the minimum of (1) the highest trait value with a standard error less than or equal to ten times the average minimum standard error, and (2) the highest trait value obtained in the distribution. It was desired to have the floor for a test level be less than or equal to the floor for any higher level. When disordinal floors were obtained, both floors were set equal to the lower of the two floors. Ordinal ceilings were produced by a comparable procedure. The ceiling for examinees with perfect scores was set at a tenth of a LOGIST trait unit above the ceiling for examinees without perfect scores.

Scoring

Two scoring programs were developed: number-right scoring and item-response scoring.

The number-right scoring procedure produced maximum likelihood trait estimates based on number-right scores. The compound-binomial likelihood function was approximated using a procedure described in Lord & Novick (1968, p.525). A number-right score to trait estimate table was produced.

The second scoring procedure produced maximum likelihood trait estimates based on patterns of item responses. Lord (1974; 1980, pp.179-181) has described such estimation and provided LOGIST to obtain these trait estimates. One disadvantage of using LOGIST is the amount of computer time needed to obtain trait estimates. Because of the large amount of scoring to be done, a more efficient scoring program was written in Assembler. The item-response scoring program was an iterative procedure that found a trait estimate that maximized the likelihood of the observed pattern of right and wrong answers. In this program the algorithm used for maximizing the likelihood was modified from that used in LOGIST; the modification estimated traits more rapidly. Also, the number-right trait estimate was used as a starting value in the iterative procedure, which differs from LOGIST's starting value. (LOGIST's starting value is the number-right adjusted for omits and then standardized.) Using trial data, the estimation procedure was found to converge in almost all cases; exceptions were some examinees (1 to 5 per cent, depending on the test) with low scores who had

diverging, low trait estimates. When the iterative procedure did not converge after ten iterations, the examinee's trait estimate was set equal to the starting value, i.e., the number-right trait estimate.'

Figure 2 contains a sample scatter plot of the number-right and item-

Insert Figure 2 about here

response scale scores. Further research showed the number-right and item-response trait estimates were highly correlated, and that the number-right trait estimates were unbiased estimates of the item-response trait estimates throughout the range of scores. Consistent with information theory's predictions, the number-right trait estimates had higher observed score standard deviations, higher standard errors of measurement, and lower split-half correlations than the item-response trait estimates. Also, the relationship between number-right and item-response trait estimates was found to be very similar in the Black and the Other ethnic groups. In particular, it was clear that low scoring groups, regardless of ethnic group, were not "penalized" by item-response scoring.

Content area total and battery total scores were taken to be averages of the trait estimates contributing to the total. The different content area scales had been chosen to minimize variability in the standard deviations of the scales, because these standard deviations affect the implicit weighting in obtaining area totals.

Norms

Two sets of normative distributions were produced using the two scoring procedures: item-response trait estimates and number-right trait estimates. Detailed comparisons between these distributions are now being made to see if one set of norms would be sufficient.

In scoring the tests for the norms (and for all subsequent user scoring) omitted items and not reached items were scored as wrong answers. The decision to treat omitted and not reached items as wrong answers was based on two primary considerations: (1) If a student is not penalized for not reached items and this fact is known to teachers and students, it becomes possible to manipulate scores through test directions and changes in test-taking strategies. (2) Unless omitted and not reached items are treated as wrong answers, the hand scoring of tests becomes complicated.

Scoring Reports

Numerous scoring reports are available for CTBS/U and TCS. Several innovations in these reports were made as the result of the use of IRT:

1. The choice between number-right and item-response scale scores is provided.
2. Confidence intervals for scale scores (trait estimates) are variable depending on the examinee's level of performance and the type of scoring (number-right or item-response) used. These confidence intervals are based on standard errors of measurement available from information theory (Lord, 1980).
3. A method of producing and reporting objective mastery scores using Bayesian and IRT procedures has been adopted. This method is described in the Appendix.

In brief, the objective mastery score procedure first checks to see if an examinee has performed similarly across the objectives in the test. If there is consistent performance, reflecting unidimensionality, information about the examinee's performance on the remainder of the test

is combined with the actual performance on the objective in obtaining a mastery score. The weights used in this combination are the result of Bayesian and IRT considerations. If an examinee performs differentially across the objectives in a test, indicating multidimensionality, data from the remainder of the test are not pooled, and the actual performance on the objective is reported as the mastery score. The Bayesian/IRT procedure produces smaller standard errors for mastery scores than are available from the raw mastery scores. The last paragraph of the Appendix discusses a sample objective mastery scoring report.

Locator Test

The use of IRT will facilitate the development of the CTBS/U Locator Test. There are two levels of the locator and each level has 20 reading vocabulary and 20 mathematics computation items. A minimum of 1,000 examinees took one level of the Locator Test and the appropriate level of CTBS/U. In each content area for each CTBS/U level, the locator and CTBS/U items will be analyzed simultaneously using LOGIST. Using the means and standard deviations of the CTBS/U b_1 parameters in standardization (scale score) units and in LOGIST units, the Locator Test item parameters will be transformed to scale score units. A number-right scoring table in scale score units will be obtained for the Locator Test. Using the standard error curves for the CTBS/U tests, a number-right locator score can be directly related to the CTBS/U test level that would most likely produce the lowest

standard error of measurement for that examinee.

Discussion

Evaluation of IRT in the Item Tryout

Most of the CTB/McGraw-Hill staff involved in the IRT item tryout had been involved in several previous item tryouts that used traditional item statistics. The consensus was that the IRT procedure produced a clearer picture than the traditional procedure of the relative contribution of each item to the test's statistical qualities, particularly in terms of measurement accuracy. The editorial staff readily adapted to the description of item difficulty in terms of the scale score where the item measured best; however, there was some confusion about the relationship between an item's point biserial and the item's discrimination. It seemed that part of this confusion arose because of the interaction of difficulty and discrimination in determining an item's point biserial. Figure 3 shows

Insert Figure 3 about here

a sample plot of point biserials and discriminations; the two variables are highly related except when the p-value was less than or equal to .27.

(Of course, such difficult items are rarely used, regardless of their discrimination.) Additional confusion arose from differences between the discrimination rating (Good, Fair, Poor) and words (Acceptable and Unacceptable) that had been used in the past to describe point-biserial values (see Figure 3). The confusion was resolved by more detailed discussions between editorial and psychometric staff on the relationship between discrimination and point biserials.

It was noted earlier that in the tryout phase, c-values were not estimated but were fixed at $1/A_1 - .05$. This procedure worked except in one case when one item in a test had a p-value less than $1/A_1 - .05$. In this case, LOGIST set the c-values for all the items in the test equal to this low p-value. In order to get reasonable c-values in such a situation, it is necessary to exclude items with p-values less than $1/A_1 - .05$.

The Automatic Test Selection program gave the editorial staff a good start at picking a test. Two types of Automatic Test Selection runs were available: no-float and float runs. The no-float runs allowed items to be selected only from those tried out at that nominal level. The float runs allowed items to be selected from items tried out at that nominal level as well as from those tried out at one level above or below the nominal level. The float runs were much more useful than the no-float runs in choosing tests, and after a while the editorial staff ceased to review or request no-float runs.

It would seem from a theoretical point of view that an even better Automatic Test Selection procedure would make eligible any item with appropriate difficulty and content, even if that item had been tried out two or more levels away. However, the more extreme an item is in terms of difficulty, the greater the standard error of the item's difficulty (in scale score units) and discrimination. If such items were chosen, there would be less confidence that the tryout item statistics would accurately describe later performance of the item. Also, items with unexpectedly extreme difficulties probably are measuring a trait or traits other than the one intended. This unpredictability of the item would make it a poor choice.

Recall that an item could be chosen for a test if its difficulty (in scale score units) fell in subranges 1 to 6. In some cases excellent items had difficulties just outside the acceptable ranges. The Automatic Test Selection would be improved by permitting such items to be chosen; instead of satisfying subrange quotas, a better procedure would select items to minimize some function of the standard error curve within some specified range of scale scores. However, the subrange quota procedure was very useful in that editorial staff could readily understand and implement it.

It appeared that items that did poorly on the basis of one criterion (for example, bias) frequently seemed to do poorly on the basis of other criteria (fit or discrimination). This is a general impression, and objective research about relationships among the item quality criteria has not been conducted. Also, frequently the poor statistical performance of an item could be satisfactorily explained by an examination of the item content. However, there certainly were statistically poor items that appeared to be flawless in terms of content and format.

The items identified as ethnically biased tended to be scattered in terms of content and format, and frequently no rational explanation was suggested for the way ethnic groups performed differently on them. There was a group of coin identification items in the Mathematics Concepts and Applications test at Level A which appeared "biased" against the Other group. It seemed reasonable that Black and Other groups could have had differential experiences handling money. The content of the items was clearly appropriate, and the items were not removed from the test.

In choosing items to maximize the various criteria, "capitalization on chance" no doubt occurred, and the test statistics based on the tryout data could look better than the statistics based on the "cross validation" standardization data. The larger the tryout sample size, the less such an effect would be expected. In general the sample sizes were felt to be quite adequate to estimate difficulty, discrimination, and fit. The bias analyses not only required the estimation of item difficulties but also the estimation of the difference between the item difficulties for the two ethnic groups. Difference scores are notoriously unreliable. While the bias procedures should have identified items with very different performance in the two ethnic groups, the rank ordering of the items in terms of more subtle differences in performance was probably not as systematic as it would have been if larger minority group samples could have been obtained.

In general, the IRT item tryout procedure was found to be very powerful. With a larger item pool the procedure would be even more useful. Systematic, detailed comparisons of tryout predictions of test performance with standardization test performance are now being planned.

Evaluation of IRT in the Standardization

The most unexpected results from the use of IRT occurred in the standardization phase. Different interlevel equating procedures were used for TCS and CTBS/U, and surprises came with each procedure.

TCS equating. As described earlier, the trait estimates were to be used to equate adjacent test levels administered to matched samples. It was expected that this would essentially be a linear equipercentile equating. Trait estimates were to be obtained in each sample, and trait values at the same percentile in the two groups were to be plotted. A straight line would then be fitted to the plot. The fact that the floors and ceilings of the tests were set rationally (but arbitrarily) would not influence the bulk of the scatter plot or the constants of the linear equating. The surprise was that the equipercentile plots were not linear. The most curvilinear plot appears in Figure 4.

Insert Figure 4 about here

Further thought led to the idea that the nonlinearity of the plots was the logical result of the fact that the test levels had different standard error curves. The observed score distributions had different shapes because of differences in the location of the different amounts of error variance. A series of simulations replicated the curvilinear plots. The simulations also indicated that the test levels could be adequately equated (made tau-equivalent, Lord & Novick, 1968) by equating the means and standard deviations of the trait estimates for the matched samples after setting reasonable, but arbitrary, floors and ceilings. More information about this topic is available in Yen (1982).

CTBS/U equating. The CTBS/U interlevel equating involved the use of the linkage books and the equating of estimated item parameters for the same items appearing in different books. The two sets of item parameter estimates to be equated were plotted and correlated. The parameters tended to be very highly correlated (i.e., in the .80's and .90's for the a parameters and in the .90's for the b parameters). The scatter plots were very linear. Occasionally, there were outliers, which usually were items with extreme b values and low a values; the outliers were deleted from the calculation of the equating constants. Two clear context effects were found. In one, the sample item (for which the correct answer is provided by the test administrator) in a Language Mechanics test in the D-E linkage book was essentially the same as the third item in the book. In the Level D book the sample item was quite different from the item in question. The item was much easier in the linkage book than in the Level D book. This item was deleted from the equating.

Another clear context effect was found with some Reading Comprehension items involving the category objectives "The student will differentiate between reality and fantasy ..." and "The student will recognize techniques of ... figurative speech." For items measuring the first objective, examinees identified statements that could not really happen, such as "The wind begged us to open the door." For the second objective, the examinees interpreted figurative writing, such as "You move like a turtle." In one book items measuring the two objectives were juxtaposed; in this book the items were more difficult than in another book where the items were separated by other reading comprehension tasks. These items were deleted

from the computation of the equating constants.

The biggest surprise occurring with the CTBS/U interlevel linking related to the type of scale produced. As grade increased, the scale scores had decreasing standard deviations and corresponding decreasing standard errors of measurement and increasing item discriminations. This effect was most apparent with the Mathematics tests. Table 7 contains the standard

Insert Table 7 about here

deviations of Mathematics Computation scale scores by grade.

This result was unexpected because the scaling procedure used with previous tests, Thurstone's absolute scaling, produced a scale with standard deviations increasing with grade. Psychological reasons can be presented to explain decreasing as well as increasing standard deviations. For example, Mathematics Computation in particular can be seen as a cumulative skill. More difficult items should be more discriminating (i.e., provide more information about computation skill) because they tend to involve more complex, sequential activities. Also, one can argue that because the computation skills that are being tested are directly taught to fewer students (except for remedial reasons) as grade increases, students tend to become more homogeneous with respect to the tested computation skills.

Even though the obtained scale appeared plausible, the scaling itself was very carefully double-checked because it was different from previous results. Data and computer programs that had, of course, already been carefully checked were rescrutinized. Several psychometric hypotheses about how such a scale could be spuriously produced were also examined: (1) overestimation of the c parameter in general or especially for difficult items; (2) differential standard error of the b estimates when obtained with different samples; (3) non-linear relationships of scales for adjacent test levels; (4) systematic differences in the parameters of easier items when they were estimated alone versus when they were estimated with more difficult items; and (5) inconsistency between equating using b parameters and equating using θ values: None of these hypotheses was found to have any foundation. The data analyzed for hypothesis 5 were particularly reassuring. At grades 4, 6 and 8 matched samples were available that had taken adjacent test levels. Using the linked scale, trait estimate (scale score) means were obtained. The matched samples had very similar means, indicating accurate equatings. Further detailed examinations of the performance of the IRT test development and standardization results are now being planned.

Evaluation of IRT in Test Scoring

Recall that there were three scoring innovations produced by IRT: (1) a choice between number-right and item-response scoring; (2) variable standard errors of measurement; and (3) a Bayesian/IRT objective mastery score.

The item-response scoring appears very useful. Examinees can be identified that have the same number-right scores and quite different item-response scores (trait estimates). For example, Table 8 shows the

Insert Table 8 about here

correct answers for two such examinees who took a TCS Verbal Reasoning test. In the table the items have been reordered in terms of increasing difficulty. The most obvious explanation of examinee B's pattern of responses is that the examinee was guessing. Given what is known about the item difficulties, the item-response scale scores assigned to these examinees reflect in a reasonable manner the differences in their response patterns.

One difference between traditional and IRT results relates to the variability of the standard error of measurement over examinees who have taken the same test. Traditional procedures use one standard error of measurement for all examinees. This constant value is approximately the average of the standard errors for the individual examinees. The traditional constant standard error is very misleading because it overestimates the standard error of some examinees and grossly underestimates the standard error for other examinees, particularly those examinees performing at the floor and ceiling of the test. In multiple-choice tests where guessing occurs, measurement accuracy is especially poor at the floor of the test. Traditional procedures that use a constant

standard error mask the poor quality of measurement obtained with extreme scores.

The objective mastery score procedure is described in the Appendix. It appears to be very useful, particularly in the identification of differential performance on the objectives in a test. Such differential performance is a sign of multidimensionality for the examinee, and multidimensional performance indicates that any summary score for the examinee for the test, be it a number-right score or an item-response scale score, has limited usefulness. The wide confidence intervals (see Figure A) in these multidimensional situations appear very appropriate and could provide information helpful to a teacher. The ultimate evaluation of the usefulness of the objective mastery scores must await their use by teachers.

Overall Evaluation of IRT

It was a tremendous amount of work to change from traditional test development and standardization procedures to IRT procedures. The change involved not only procedural changes but also extensive rethinking of psychometric issues. IRT applications are relatively new, and many questions about them have required research and careful thought. Additional research and trial-and-error are needed. It is clear to the author that the change to IRT was worth it and has paid off already in terms of the clarification of psychometric issues, particularly those related to equating, the clarity IRT brings to test development, and in the improvement in the scoring process. IRT has tremendous potential for equating, item pooling, tailored

testing, and score interpretation, and it deserves an important role in the future of testing.

References

- Allen, M., & Yen, W.M. Introduction to measurement theory. Monterey, CA: Brooks/Cole, 1979.
- Bloom, B.S. (Ed.) Taxonomy of educational objectives, the classification of educational goals, Handbook I: Cognitive domain. NY: David McKay Co., Inc., 1956.
- California Achievement Tests, Form C. Monterey, CA: CTB/McGraw-Hill, 1977.
- Collected Algorithms from Communications of the Association for Computing Machinery. NY: Association for Computing Machinery, 1971.
- Comprehensive Tests of Basic Skills, Form U. Monterey, CA: CTB/McGraw-Hill, 1981.
- Lord, F.M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.
- Lord, F.M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Laurence Erlbaum Associates, 1980.
- Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Menlo Park, CA: Addison-Wesley, 1968.
- Novick, M.R., & Jackson, P.H. Statistical methods for educational and psychological research. NY: McGraw-Hill, 1974.
- Test of Cognitive Skills. Monterey, CA: CTB/McGraw-Hill, 1981.
- Test of Cognitive Skills Test Coordinator's Handbook and Guide to Interpretation, Preliminary Edition. Monterey, CA: CTB/McGraw-Hill, 1981.

Wood, R.L., Wingersky, M.S., & Lord, F.M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (RM-76-6).

Princeton, NJ: Educational Testing Service, 1976.

Yen, W.M. Obtaining some degree of correspondence between unequatable scores: A comparison of item response theory and equipercentile equating methods. Paper presented at the American Educational Research Association meetings, New York, March 1982.

Acknowledgements

Because of the size of this test development project, many people at CTB/McGraw-Hill were involved in one or more phases of it. It is not possible to acknowledge them all here, but I would like to thank George R. Burket and Donald Ross Green for their extremely valuable counsel during the conceptualization of the IRT aspects of this project. Of course, this work would never have been done without the theoretical foundation for IRT created by Frederic M. Lord.

Author's Address

Please address correspondence concerning this paper to Wendy M. Yen, CTB/McGraw-Hill, Del Monte Research Park, Monterey, CA 93940.

Table 1

Number of Items by Test and Level for
CTBS/U

Test	Level									
	A	B	C	D	E	F	G	H	J	
Visual Recognition	18									
Sound Recognition	17									
Oral Comprehension	15	15								
Word Attack		32	30	40	35					
Vocabulary	16	17	25	25	30	45	45	45	45	
Reading Comprehension			25	25	30	45	45	45	45	
Spelling				25	25	30	30	30	30	
Language Mechanics				20	20	30	30	30	30	
Language Expression		17	20	25	25	45	45	45	45	
Mathematics Computation			20	20	36	40	40	40	40	
Mathematics Concepts and Applications	15	20	25	30	40	45	45	45	45	
Reference Skills						20	20	20	20	
Science				25	30	40	40	40	40	
Social Studies				25	30	40	40	40	40	

Table 2
Item Overall Quality

Overall Quality	Rating		
	Discrimination	Fit	Bias *
1	G	G	G
2	G	G	F
3	G	F	G
4	G	F	F
5	F	G	G
6	F	G	F
7	F	F	G
8	F	F	F
9	G	G	P
10	G	G	VP
11	G	F	P
12	G	F	VP
13	F	G	P
14	F	G	VP
15	F	F	P
16	F	F	VP
17	G	P	G
18	G	P	F
19	F	P	G
20	F	P	F
21	P	G	G
22	P	G	F
23	P	F	G
24	P	F	F
25	G	P	P
26	G	P	VP
27	F	P	P
28	F	P	VP
29	P	G	P
30	P	G	VP
31	P	F	P
32	P	F	VP
33	P	P	G
34	P	P	F
35	P	P	P
36	P	P	VP
37**	anything	anything	ND
or	ND	ND	anything

Note: G = Good, F = Fair, P = Poor, VP = Very Poor, ND = No Data.

* For CTBS/U the Black-Other bias rating was used; for TCS the composite bias rating was used.

** If there were inadequate data for estimating its item parameters, an item received No Data discrimination and fit ratings; the criteria for the No Data bias rating are described in the "Bias Analysis" section.

Table 3
CTBS/U Standardization Sample Sizes

Level	Fall		Spring	
	Grade	Sample Size	Grade	Sample Size
A	K.1	7123		
B			K.8	4769
	1.1	4393		
C			1.8	8070
D	2.1	4393		
			2.8	7348
E	3.1	4540		
			3.8	7671
F	4.1	3890		
			4.8	3913
G			4.8	3435
	5.1	4254		
			5.8	11835
	6.1	5063	6.8	4527
H			6.8	4124
	7.1	7652	7.8	18992
	8.1	7947	8.8	6825
J			8.8	5636
	9.1	3119		
			9.8	7810
	10.1	3094	10.8	5972
	11.1	3037	11.8	5494
	12.1	3028	12.8	4579

Table 4
TCS Standardization Sample Sizes

Level	Grade	Sample Size
1	2.1	7806
	3.1	4029
2	3.1	4147
	4.1	7935
	5.1	4335
3	5.1	3948
	6.1	9466
	7.1	6461
4	7.1	6869
	8.1	13947
	9.1	2689
5	9.1	2915
	10.1	2866
	11.1	2813
	12.1	2812

Table 5
Number of Items in the
CTBS/U Linkage Books

Test	Book							
	A-B	B-C	C-D	D-E	E-F	F-G	G-H	H-J
Oral Comprehension	20							
Word Attack		20	30	40				
Vocabulary	20	20	25	25	30	45	45	45
Reading Comprehension			25	25	30	45	45	45
Spelling				25	25	30	30	30
Language Mechanics				20	20	30	30	30
Language Expression		20	20	25	25	45	45	45
Mathematics Computation			20	20	36	40	40	40
Mathematics Concepts and Applications	20	20	25	30	40	45	45	45
Reference Skills						20	20	20
Science				25	30	40	40	40
Social Studies				25	30	40	40	40

Table 6

LOGIST Runs made for Linking CTBS/U Levels

	Book	Level	Items		
			In Shelf Only	In Link Only	In Common
Run 1	{	Shelf A	Responses	NR	Responses
		Link A-B	NR	Responses	Responses
Run 2	{	Link A-B	NR	Responses	Responses
		Shelf B	Responses	NR	Responses
Run 3	{	Shelf B	Responses	NR	Responses
		Link B-C	NR	Responses	Responses
etc.					

Note. "NR" means "not reaches" were used in place of item responses.

"Responses" means the item responses were used.

Table 7
Standard Deviation of
Mathematics Computation Scale
Scores by Grade

Grade	St. Dev.
2.1	99
3.1	50
4.1	38
5.1	25
6.1	24
7.1	19
8.1	21
9.1	18
10.1	18
11.1	19
12.1	22

Table 8

Example of Two Examinees with the
Same Number-Right Score and
Different Item-Response Scale Scores

Item Difficulties	Examinee	
	A	B
305	+	
325	+	
334	+	
361	+	+
397		
398	+	
398	+	+
409		+
410		+
423		
423		
425		
426		+
438		
446		
456		
457		+
461		
464		
490		

Note: "+" indicates a correct answer. Examinee A had an item-response score of 373; Examinee B had an item-response score of 213, which was the lowest possible score for this test.

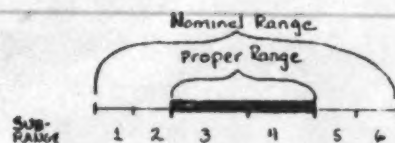
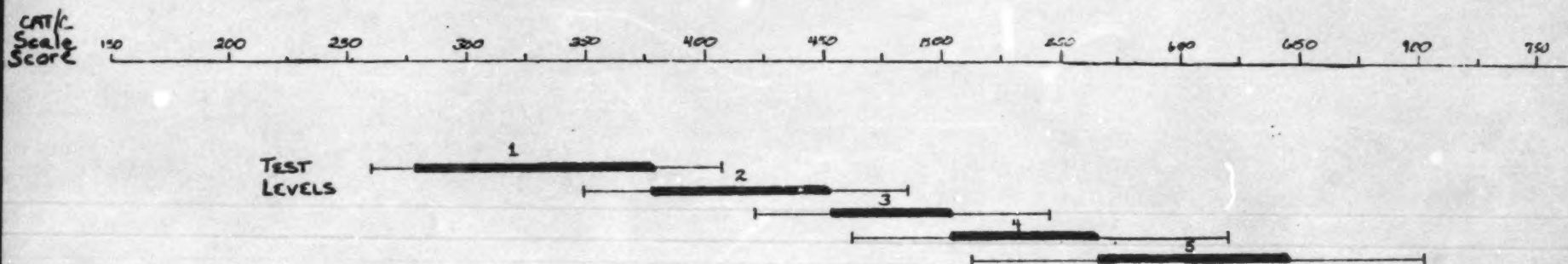


FIGURE 1. NOMINAL AND PROPER RANGES FOR TCS

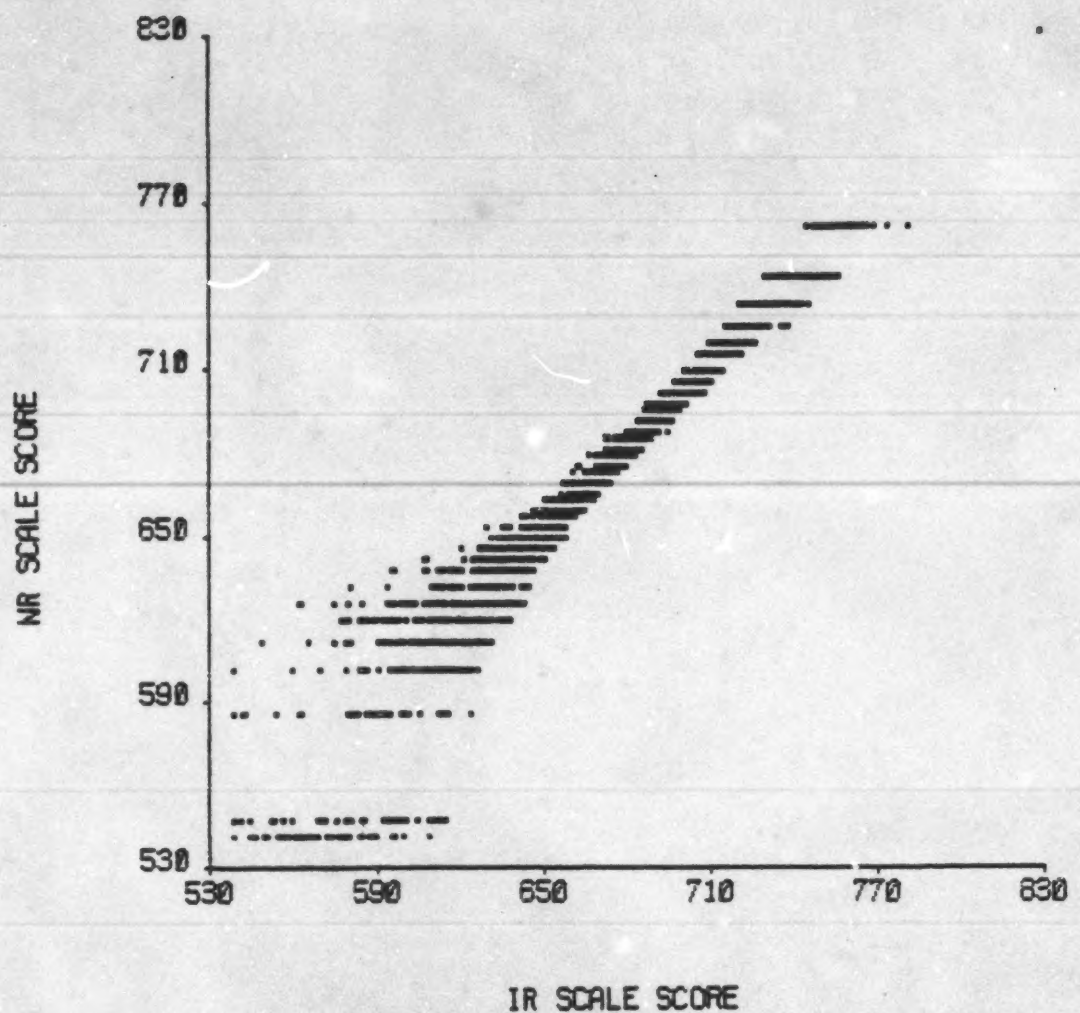


Figure 2. Relationship between number-right (NR) and item-response (IR) scale scores for CTBS/U level G Reading Vocabulary at grade 6.

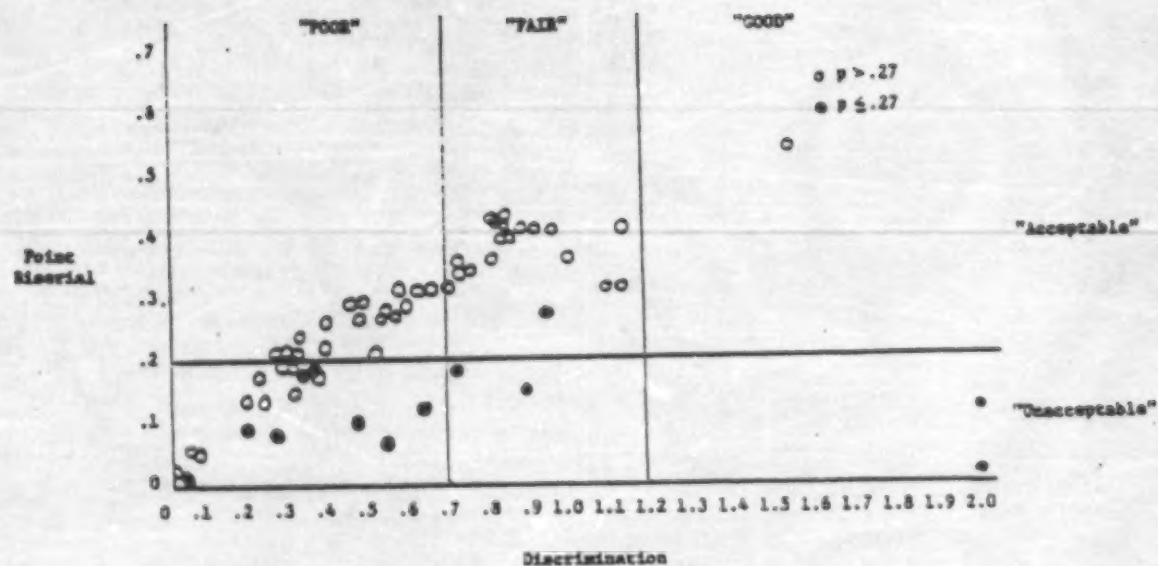


Figure 3. Relationship between point biserials and discriminations for TCS Verbal Reasoning Level 4 "Other" tryout sample ($n = 363$).

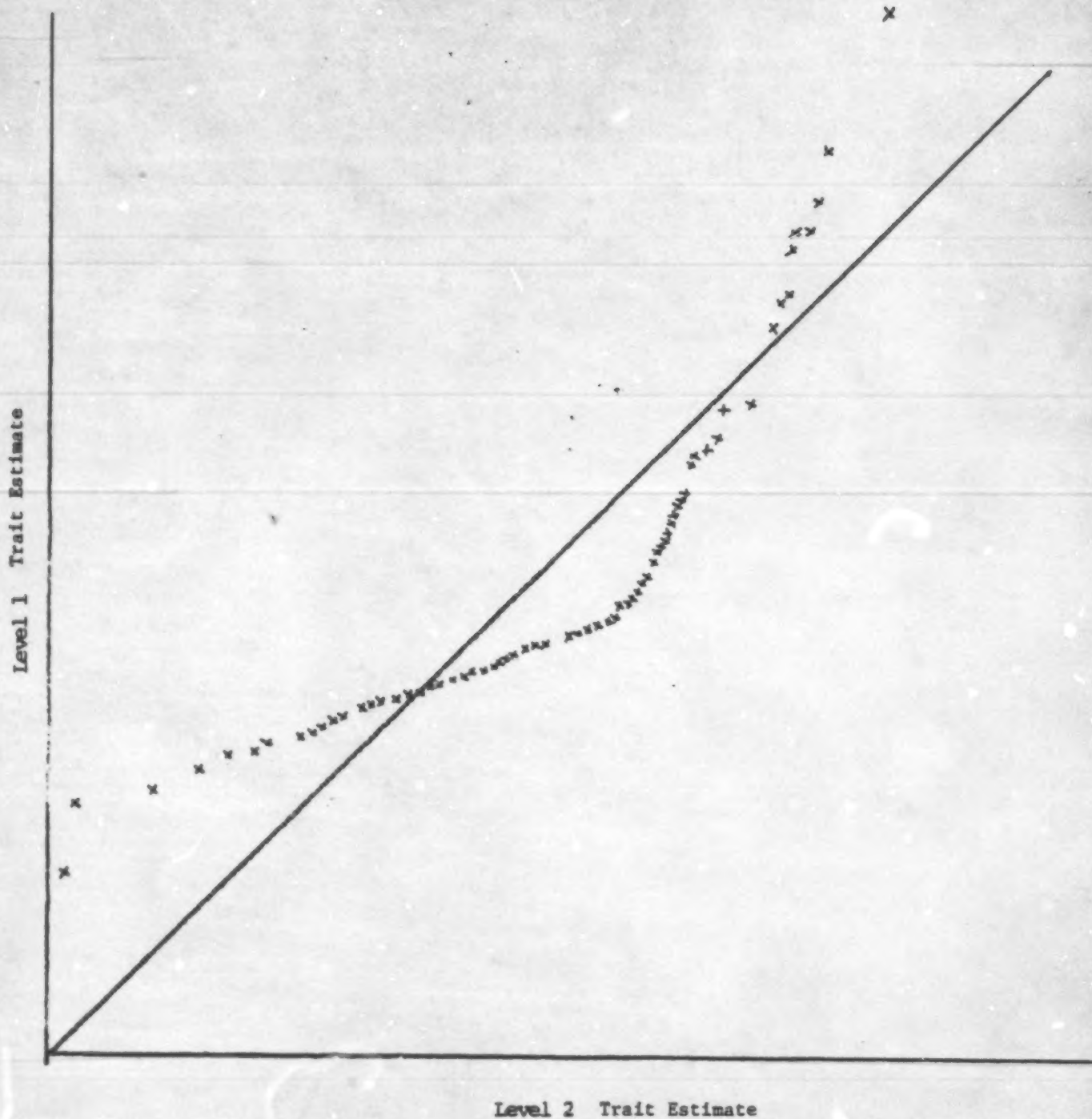


Figure 4. Equipercentile plot for TCS Verbal Reasoning levels 1 & 2.

Appendix: A Bayesian/IRT Objective Mastery Score Procedure

The CTBS/U tests are composed of category objectives with from 4 to 20 items per objective. It was desired to report mastery scores, or estimated true scores, for the objectives and confidence intervals for these true scores. Because of the small number of items associated with most objectives, these confidence intervals could be very wide. It was desired to take into account information available from the examinee's related performance on the rest of the items in the test and use this information to decrease the standard error of the true score estimate. The combination Bayesian/IRT procedure that was developed to produce objective mastery scores is now described.

The prior true score estimate for objective j is

$$\hat{T}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}(\hat{\theta})$$

where $\hat{\theta}$ is the trait estimate (scale score) found from the overall test performance, and there are n_j items in the objective. The objective mastery score, or mean of the posterior distribution, is

$$\hat{\eta}_j = w_j \hat{T}_j + (1 - w_j) \frac{x_j}{n_j},$$

where x_j is the observed number-right score on objective j and w_j is the weight given to the prior.

From Novick and Jackson²(1974,p.113),

$$\hat{T}_j = \frac{a}{a+b},$$

and the prior variance is

$$\sigma^2(\hat{T}_j) = \frac{ab}{(a+b)^2(a+b+1)},$$

where a and b are the parameters of the Beta distribution. (Note that here the a and b follow Novick and Jackson's notation and are not the item discrimination and difficulty.)

The posterior mean is

$$\hat{\eta}_j = \frac{p}{p+q},$$

and the posterior variance is

$$\sigma^2(\hat{\eta}_j) = \frac{pq}{(p+q)^2(p+q+1)},$$

with

$$p = a + x_j$$

and

$$q = b + n_j - x_j.$$

Solving for a and b ,

$$a = \hat{T}_j \left[\frac{\hat{T}_j(1 - \hat{T}_j)}{\sigma^2(\hat{T}_j)} - 1 \right];$$

$$b = (1 - \hat{T}_j) \left[\frac{\hat{T}_j(1 - \hat{T}_j)}{\sigma^2(\hat{T}_j)} - 1 \right].$$

Solving for p and q,

$$p = \hat{\tau}_j \left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right] + x_j ;$$

$$q = (1 - \hat{\tau}_j) \left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right] + n_j - x_j .$$

Thus, the posterior mean is

$$\hat{\eta}_j = \frac{\left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right]}{\left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right] + n_j} [T_j] + \frac{n_j}{\left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right] + n_j} \left[\frac{x_j}{n_j} \right] ,$$

and

$$w_j = \frac{\left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right]}{\left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right] + n_j} .$$

The posterior variance is

$$\sigma^2(\hat{\eta}_j) = \frac{\hat{\eta}_j(1 - \hat{\eta}_j)}{\left[\frac{\hat{\tau}_j (1 - \hat{\tau}_j)}{\sigma^2 (\hat{\tau}_j)} - 1 \right] + n_j + 1} .$$

IRT can be used to estimate $\sigma^2 (\hat{\tau}_j)$.

$$\sigma^2 (\hat{\tau}_j) \doteq 1/I(T_j, \hat{\tau}_j) .$$

$I(T_j, \hat{\tau}_j)$ is the information that $\hat{\tau}_j$ contributes about T_j . From Lord (1980,p.85)

we know that

$$\begin{aligned}
 I(T_j, \hat{T}_j) &= \frac{I(\theta, \hat{T}_j)}{\left[\frac{dT_j}{d\theta} \right]^2} \\
 \frac{dT_j}{d\theta} &= \frac{d \left[\frac{1}{n_j} \sum_{i=1}^{n_j} P_{ij}(\theta) \right]}{d\theta} \\
 &= \frac{1}{n_j} \sum_{i=1}^{n_j} \frac{dP_{ij}(\theta)}{d\theta} \\
 &= \frac{1}{n_j} \sum_{i=1}^{n_j} P'_{ij}(\theta).
 \end{aligned}$$

From Lord (1980, p.79), we also know that

$$\begin{aligned}
 I(\theta, \hat{T}_j) &\doteq I(\theta, \hat{\theta}), \\
 \text{where} \quad I(\theta, \hat{\theta}) &= \sum_{j=1}^J \sum_{i=1}^{n_j} \frac{[P'_{ij}(\theta)]^2}{[P_{ij}(\theta)][1 - P_{ij}(\theta)]},
 \end{aligned}$$

and there are J objectives in the test. Then,

$$I(T_j, \hat{T}_j) \doteq \frac{I(\theta, \hat{\theta})}{\left[\frac{1}{n_j} \sum_{i=1}^{n_j} P'_{ij}(\theta) \right]^2}.$$

For each objective, the values of p and q are referred to tables of the Beta distribution to obtain a 17th percentile to 83th percentile confidence (or credibility) interval for $\hat{\eta}_j$.

For some examinees, pooling information from the remainder of the test is not helpful because the examinee performed quite differently on the different objectives in the test. The following statistic was used to identify examinees with differential performance on the objectives in a test:

$$Q = \sum_{j=1}^J \frac{n_j \left(\frac{x_j}{n_j} - \hat{T}_j \right)^2}{\hat{T}_j (1 - \hat{T}_j)}.$$

If Q was less than or equal to the chi-square value at the .10 level for J degrees of freedom, $\hat{\eta}_j$ was used for all J objectives. Otherwise, the objective scores were taken to be the observed proportions passing the objectives, $\hat{\eta}_j = \frac{x_j}{n_j}$, with confidence intervals based on $p = x_j$ and $q = n_j - x_j$.

Figure A displays a sample scoring report using this method. For Language Expression and Social Studies the observed proportions passing were used instead

Insert Figure A about here

of the weighted objective mastery scores, and the confidence intervals were correspondingly wider than those for the other tests. Within Language Expression and within Social Studies, the examinee performed very differently on the different objectives. Within each of the other tests, the examinee performed similarly across the objectives in the test.



Figure A. Sample objective mastery score report with confidence intervals for estimated true scores, $\hat{\eta}_j$.